Journal of Network and Computer Applications 38 (2014) 125–134

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Balancing trajectory privacy and data utility using a personalized anonymization model

Sheng Gao *, Jianfeng Ma, Cong Sun, Xinghua Li

School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China

ARTICLE INFO

Article history:
Received 11 July 2012
Received in revised form
1 January 2013
Accepted 18 March 2013
Available online 29 March 2013

Keywords:
Trajectory *k*-anonymity set
Trajectory similarity and direction
Trajectory distance
A personalized anonymization model

ABSTRACT

With the widespread use of location-based services (LBS), the number of trajectories gathered by location service providers is dynamically growing. On the one hand, mining and analyzing these spatiotemporal trajectories can help to work out a mobile-related strategic planning; on the other hand, knowledge of each trajectory can be used by adversaries to identify the user's sensitive information and lead to an unpredictable harm. The concept of trajectory *k*-anonymity extends from location *k*-anonymity that has been widely used to address this issue. The main challenge of trajectory *k*-anonymity is the selection of a trajectory *k*-anonymity set. However, existing anonymity methods ignore the trajectory similarity and direction, assuming that it has little impact on privacy. Thus, it cannot provide a preferable trajectory *k*-anonymity set. In this paper, we propose to use trajectory angle to evaluate trajectory similarity and direction, and construct an anonymity region on the basis of trajectory distance. Considering the various preference settings on the proportion of trajectory privacy and data utility in different scenarios, we propose a personalized anonymization model to select the trajectory *k*-anonymity set. Experiment results prove that our method can provide an effective trajectory *k*-anonymity set under various proportions of trajectory privacy and data utility requirements, while the efficiency just reduces a little.

1. Introduction

The explosion of mobile devices equipped with powerful wireless communication ability, together with the rapid progress of mobile positioning techniques such as global positioning systems (GPS), radio frequency identification (RFID) and so forth, have greatly facilitated the prosperity of location-based services. While users sharing the mobile services, a large number of trajectories might be collected by service providers. Mining and analyzing these spatiotemporal trajectories (Ivanov, 2012) can help people to make a mobile-related decision, for instance, merchants can decide the place where to build a restaurant or a supermarket by analyzing trajectories of customers in a certain area (Cao et al., 2010) and tourism company can make a travel recommendation schedule by monitoring trajectories of visitors in a city (Zheng et al., 2009). For such practical applications, the main step is to explore accurate and applicable knowledge, which is out of the scope of this work.

The publication of spatiotemporal trajectories is a double-edged sword. Although mining trajectories can bring many advantages to multiple commercial applications, the disclosure of those spatiotemporal information contained in trajectories may threaten individuals sensitive information, such as home addresses, travel habits, political beliefs, health conditions, personal interests, and so on. To cope with the problem, trajectory *k*-anonymity is presented to anonymize *k* trajectories at least over a time span (Nergiz et al., 2008; Abul et al., 2008; Xu and Cai, 2008; Yarovoy et al., 2009). It is an extension of location *k*-anonymity (Gruteser and Grunwald, 2003), which conceals a user's trajectory with the assistance of the other $k-1$ trajectories at least. Instead of revealing the exact trajectory of a user, an obscure path called *anonymized path* that contains at least *k* trajectories is reported.

To ensure a high quality of anonymization, the main challenge is to determine a trajectory *k*-anonymity set. We observe that the selection of anonymity trajectories affects the trajectory privacy protection level and data utility. Shin et al. (2010a) noticed that existing location *k*-anonymity model regarded the location as sole information when achieving anonymization. However, the disclosure of users' movement directions can cause adversaries to identify a mobile user who submitted a LBS request. They proposed to improve the location *k*-anonymity model by taking a user's direction of movement into account during the anonymous request process. To best of our knowledge, most of trajectory *k*-anonymity methods anonymize the trajectories without taking the similarities and directions among them into consideration. As Shin et al.

* Corresponding author. Tel.: +86 15991720644.
E-mail address: sgao555@gmail.com (S. Gao).

1084-8045/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2013.03.010

(2010a) demonstrated, the trajectory directions affect the trajectory privacy protection level. However, the differences among trajectories may also affect the quality of anonymization. The trajectories can be identified easily with high individual differences. Meanwhile, the data utility of trajectory may reduce with the expansion of anonymity region. In the follow-up method, they also proposed to use optimal trajectory division (Shin et al., 2010b) to strengthen privacy protection and improve the quality of service (QoS). Specifically, through the partition of trajectories with minimum area of anonymity region, the privacy level was increased for the unlinkability over time and the overall quality of service was improved for the smaller anonymous regions.

Motivated by this, in this paper, we take these factors into account. In location privacy protection, Gedik and Liu (2005) proposed a privacy framework on the basis of the requirements of location privacy k and QoS from the perspective of each user and then presented a cloaking algorithm *CliqueCloak* to produce an undirected graph for location privacy protection. However, it only works with a small value of k and fails when the value of k is large. To tackle this defect, Xiao et al. (2007) improved the cloaking algorithm for a robust anonymity while considering both location privacy and QoS. In trajectory privacy protection, trajectory similarity is an important factor for trajectory clustering and anonymization. Pelekis et al. (2007) presented a framework to address the trajectory similarity search problem. The authors transformed this issue into different kinds of similarity queries according to the trajectory characteristics. Moreover, the other works proposed some typical measures for trajectory similarity including Euclidean distance (Abul et al., 2008; Huo et al., 2011; You et al., 2007), edit distance (Chen et al., 2005) and linear spatio-temporal distance (Tiakas et al., 2009). However, in some cases, all of these could not reflect the factor of trajectory similarity and direction very well. To best of our knowledge, in this paper, we first propose to use trajectory angle to evaluate trajectory similarity and direction and construct an anonymity region based on trajectory distance. We construct a personalized anonymization model to balance trajectory privacy and data utility and then translate the selection of a trajectory k-anonymity set into a constrained minimum spanning tree problem. The proportion of trajectory privacy and data utility decided by a user is dependent on the application scenario. Considering that in different application scenarios, the various preference settings on the proportion of trajectory privacy and data utility may affect the selection of trajectory k-anonymity sets, we analyze the actual privacy level and data utility under these different trajectory k-anonymity sets.

In this paper, the main contributions of our work are summarized as follows:

- We propose a personalized anonymization model with taking trajectory privacy and data utility into consideration. In particular, we consider the factors of trajectory similarity and direction for privacy protection and trajectory distance for data utility.
- We transform the optimal k trajectories selection to a constrained minimum spanning tree problem and use Greedy strategy to find an approximate optimal trajectory k-anonymity set in the trajectory graph model we constructed. The weights model the relations between trajectories under various proportions of trajectory privacy and data utility.
- We run a set of evaluations on synthetic dataset. Experiment results prove that our method can provide an effective trajectory k-anonymity set under various proportions of trajectory privacy and data utility requirements, while the efficiency just reduces a little.

The remainder of this paper is organized as follows. Section 2 summarizes related work. Section 3 introduces some basic notions and states the problem of tradeoff between trajectory privacy protection and data utility. In Section 4, we present the personalized trajectory anonymization model and discuss the function of each component in detail. Section 5 discusses the metric in terms of trajectory privacy and data utility. In Section 6, we run and analyze a set of simulations on synthetic dataset to evaluate the selection of a trajectory k-anonymity set under various proportions of requirements on trajectory privacy and data utility, and then compare the effectiveness and efficiency with the previous work. Finally, we conclude this paper and present the future work in Section 7.

2. Related work

Trajectory privacy is a special type of personal privacy, which has been concerned continuously in recent years. According to the time sequence of trajectory, existing trajectory privacy-preserving techniques can be classified into three types (Huo et al., 2011).

2.1. Dummy trajectories

Kido et al. (2005) presented two algorithms to determine the dummy trajectories for trajectory privacy protection. To be specific, the next location of a dummy is selected in a neighborhood of its current location. You et al. (2007) presented two approaches to produce consistent movement patterns in a long term. However, these methods cannot strictly ensure a good similarity between trajectories. Our previous work (Gao et al., 2012) focused on the tradeoff between location and trajectory privacy protection and QoS based on a user's partners' locations and trajectories, and then proposed a method to produce the partners' trajectories that looks like the user's trajectory.

2.2. Suppression technique

Gruteser and Liu (2004) proposed to use suppression technique to protect a user's online trajectory privacy. The sensitivity map divided areas into sensitive and insensitive according to the user's settings. Once the user entered a sensitive area, location updates were suppressed at once. Terrovitis and Mamoulis (2008) studied the privacy-preserving problem in the publication of trajectory databases. They argued that each adversary would possess different portions of users' trajectories and the data publisher was aware of the adversaries' knowledge. They proposed a method that iteratively suppressed some trajectory segments until a probabilistic constraint of disclosing whole trajectories was satisfied. However, if too many trajectory segments are suppressed, it would cause lots of information loss.

2.3. Trajectory k-anonymity technique

Trajectory k-anonymity technique that anonymizes k trajectories together is directly related to our work. As a result of the imprecision of GPS devices, Abul et al. (2008) proposed *Never Walk Alone* (*NWA*) to enforce $(k,\delta)-$*anonymity* model they presented for mobile object databases using trajectory clustering and space translation. Huo et al. (2012) improved the *NWA* by anonymizing the stay points based on *grid-based approach* and *clustering-based approach*. Domingo-Ferrer and Trujillo-Rasua (2012) proposed two heuristic methods to anonymize trajectories. One of them aims at trajectory k-anonymity by microaggregation and the other is to achieve location k-diversity while considering the reachability constraints. A new distance is proposed to improve the *NWA*, which can process those trajectories without time overlap. Nergiz et al. (2008) proposed to enforce k-anonymity by grouping the

trajectories based on *log cost metric*, and then reproduce the trajectories by randomly combining sampling locations from the anonymized regions. However, the increase of privacy level may affect the accuracy of the recuperative trajectories. Xu and Cai (2008) provided a trajectory *k*-anonymity protection when mobile devices requested LBSs continuously on the move. They anonymize a user's trajectory based on the assistance of the other historical trajectories. Yarovoy et al. (2009) tackled the challenges of anonymizing mobile objects. They proposed a new notion of *k*-anonymity in the context of moving objects and formally showed that it did not lead to privacy breach. Then two methods were proposed to create the anonymous groups that provably satisfied their proposed *k*-anonymity. The most related work in Huo et al. (2011) defined the selection of trajectory *k*-anonymity set as graph partition problem and reduced the information loss by minimizing the partition cost according to the distances among trajectories. However, they ignore the trajectory similarity and direction.

3. Preliminary

We first define some basic properties of the spatio-temporal trajectory, and then state the problems.

3.1. Basic notions

Definition 3.1 (*Trajectory model Trajcevski et al., 2004*). A trajectory model is considered as a polygonal line in three-dimensional space, which can be represented as a sequence of spatiotemporal points: $T_i=\{ID_i,(x_1,y_1,t_1),(x_2,y_2,t_2),\ldots,(x_n,y_n,t_n)\}(t_1<t_2<\cdots<t_n)$, where ID_i presents the identity of the moving object. During the time interval $[t_i,t_{i+1}]$, the object is supposed to move with uniform velocity in a straight line from (x_i,y_i) to (x_{i+1},y_{i+1}).

For example, Fig. 1 illustrates two trajectories in the two-dimensional Euclidean space, where a circle represents the location of each moving object at the corresponding sampling time $(t_1,\ldots,t_8)$. We consider that all the trajectories are with fully accurate and true original locations. To depict trajectory direction, we directly utilize trajectory angle that extends slope ratio (Gao et al., 2012) further to measure the similarity of each trajectory segment, which is defined as follows.

Definition 3.2 (*Trajectory angle*). Let T_1 and T_2 be two trajectories with $n-1$ trajectory segments. The trajectory segment in time interval $[t_i,t_{i+1}]$ of $T_k(k=1,2)$ is from (x_i^k,y_i^k) to (x_{i+1}^k,y_{i+1}^k), where (x_i^k,y_i^k) represents the location of trajectory T_k at the sample time t_i.

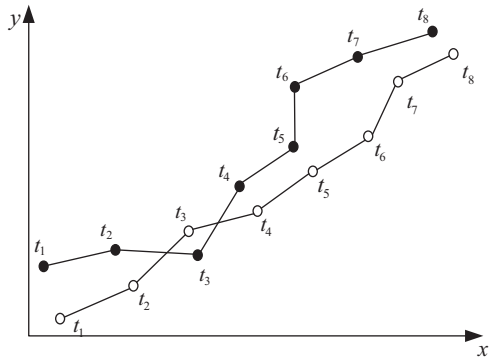

Fig. 1. The two-dimensional Euclidean space.

The trajectory segment angle θ_i ($\theta_i\in[0,\pi]$) can be calculated by (1).

$$\cos\theta_i=\frac{\overrightarrow{T_i^1}\cdot\overrightarrow{T_i^2}}{|\overrightarrow{T_i^1}||\overrightarrow{T_i^2}|}=\frac{(x_{i+1}^1-x_i^1)\cdot(x_{i+1}^2-x_i^2)+(y_{i+1}^1-y_i^1)\cdot(y_{i+1}^2-y_i^2)}{\sqrt{(x_{i+1}^1-x_i^1)^2+(y_{i+1}^1-y_i^1)^2}\sqrt{(x_{i+1}^2-x_i^2)^2+(y_{i+1}^2-y_i^2)^2}} \quad (1)$$

As $\cos x$ is a monotonically decreasing function in $[0,\pi]$, a higher value of $\cos x$ indicates a smaller angle, thus the more similar between the two trajectory segments is. Therefore, the whole trajectory similarity can be measured by (2).

$$S(T_1,T_2)=\sum_{i=1}^{n-1}\cos\theta_i \quad (2)$$

If the two trajectory segments move toward different directions, the trajectory angle θ_i ranges from $\pi/2$ to π and $\cos\theta_i\leq 0$. In this case, we set $\cos\theta_i=0$ and do not take this trajectory segment into consideration. As mentioned above, the greater value of $S(T_1,T_2)$ is, the more similar between two trajectories is.

A trajectory anonymity region is constructed based on the distances among trajectories. We first define the trajectory distance and then construct the anonymity region.

Definition 3.3 (*Trajectory distance Huo et al., 2011*). Let T_1 and T_2 be any two synchronized trajectories in time interval $[t_1,t_n]$, the distance between T_1 and T_2 is defined as the average of Euclidean distances between corresponding location samples and given by (3).

$$Distance(T_1,T_2)=\frac{\sum_{i=1}^{n}\sqrt{(x_i^2-x_i^1)^2+(y_i^2-y_i^1)^2}}{t_n-t_1} \quad (3)$$

Definition 3.4 (*Anonymity region*). Let $T_1,T_2,\ldots,T_n$ be the n trajectories. The distances among these trajectories can be computed as (4).

$$D=\{U_{i=1}^{n}D_i|D_i=Distance(T_i,T_j),j=1,2,\cdots,n\} \quad (4)$$

We select the maximum distance $D_{max}=max\{D\}$ as the diameter to construct an anonymity region R. R is formalized as a horizontal disk, denoted as $R=V(O,D_{max})$, where O represents the center of each disk and D_{max} is the diameter.

3.2. Problem statement

Trajectory *k*-anonymity is the most frequent technique that has been widely used for trajectory privacy protection. It ensures that the anonymity region R includes k trajectories at least to protect trajectory privacy. Most of the works (Abul et al., 2008; Huo et al., 2011; Domingo-Ferrer and Trujillo-Rasua, 2012) select a trajectory *k*-anonymity set on the basis of trajectory distance, which may ignore the trajectory direction and similarity. Additionally, we observe that existing trajectory similarity measurement functions such as edit distance (Chen et al., 2005), linear spatio-temporal distance (Tiakas et al., 2009) also suffer from the same defect. They cannot reflect the trajectory similarity very well. However, if we only consider the trajectory direction and similarity, the requirement of a trajectory *k*-anonymity set can result in the increase of the size of anonymity region R, which would reduce the data utility. In short, the minimum anonymity level k indicates that the trajectory cloaking should satisfy trajectory *k*-anonymity with the anonymity region R. The information loss increases with the expansion of anonymity region, which may cause a low data utility. Therefore, how to find a trajectory *k*-anonymity set that

balances trajectory privacy and data utility is the key issue, and specifically: (1) The selective k trajectories should be similar to prevent adversaries from identifying the trajectories easily; (2) The anonymity region should not be very large to prevent the data utility from reduction.

4. System overview

In this section, we introduce a personalized trajectory anonymization model and describe the function of each component in detail. We first survey the two structural modes according to the trajectory anonymity sequence, and then decide the applicable range of our anonymization model.

- *On-line mode*

The process of trajectory anonymity is done before data collection. In this mode, the collected data can be applied into analysis and application directly. Because the anonymity process is done before data collection, the data collector cannot obtain the real trajectory data. This mode can be implemented in online systems with a strong real-time process ability. However, due to the dynamic characteristic of trajectory, it is confronted with enormous challenges for trajectory privacy protection.

- *Off-line mode*

The majority of trajectory privacy protections are to anonymize trajectory after data collection. This mode is applied to off-line system, which can give high priority to consider both trajectory privacy protection and data utility. However, the trajectories need to be anonymized before reporting to data analysis and application center.

To balance trajectory privacy protection and data utility, in this paper, we propose an off-line mode of personalized trajectory anonymization model. It mainly consists of three modules depicted by Fig. 2: trajectory collection, trajectory anonymity and data analysis and application. Among of these, we mainly focus on trajectory anonymity module for trajectory privacy protection and data utility.

4.1. Trajectory pre-processing

The tasks of trajectory pre-processing phase are the same with Huo et al. (2011) that also include time span definition, trajectory equivalence class construction and trajectory synchronization.

To construct a trajectory equivalence class, Abul et al. (2008) proposed a simple algorithm $NWA_{preproc}$ that was driven by an integer parameter to process all the trajectories. However, they might prelimit the free sample time point. In this paper, we first determine the new starting time point t_p and ending time point t_q of all the trajectories. Those time points of the trajectory that do not locate between them are removed. Given n trajectories $T_1, T_2, \ldots, T_n$, each of them has a record of starting time t_{start_i} and ending time t_{end_i}, $i = 1, 2, \ldots, n$. Huo et al. (2011) determined two time intervals in advance. The trajectories whose starting time points and ending time points lay in the two time intervals are considered to be in the same equivalence class. Differ from the strategy in Huo et al. (2011), we consider that all the starting time is in a short period but much less than the ending time. Thus, we get $t_p = \max\{t_{start_i}\}$ and $t_q = \min\{t_{end_i}\}$, $i = 1, 2, \ldots, n$, where $t_p \ll t_q$. Thus, all the trajectories in the time interval $[t_p, t_q]$ construct an equivalence class, which is formalized as follows.

Definition 4.1 (*Trajectory equivalence class*). Let the new starting time point and ending time point of the n trajectories be $t_p = \max\{t_{start_i}\}$ and $t_q = \min\{t_{end_i}\}$, $i = 1, 2, \ldots, n$, respectively. All the trajectories in the time interval $[t_p, t_q]$ form an equivalence class.

Take a simple example, a time span can be set as $[9:00, 9:10]$. All the starting time and ending time of trajectories contain this time interval that are assigned in the same equivalence class. The size of time interval can be set on the basis of the extent of trajectory sparseness.

Unlike Abul et al. (2008), the sample time points of the trajectories have not been specified. We exploit trajectory synchronization function (Huo et al., 2011; Domingo-Ferrer and Trujillo-Rasua, 2012) to ensure that the trajectories in the same equivalence class have the same sample time points. For example, given any two trajectories T_r and T_s ($r \neq s$) in the same equivalence class, if a sample location in T_r has a timestamp t_e which is not in T_s, then extract a new location in T_s having the timestamp t_e.

Therefore, after finishing the tasks of trajectory pre-processing phase, those trajectories in the same equivalence class within the time span are synchronous.

4.2. Optimal trajectory graph model

In this phase, according to the user's preference settings on the proportion of trajectory privacy and data utility, we analyze the process of trajectory graph model construction, especially the weights between trajectories in detail.

Recall the requirements of trajectory k-anonymity model, the main challenge is how to find k trajectories that can provide a better privacy protection together with maximum data utility. Based on our observations, the similarity of the selected k trajectories reflects trajectory privacy protection level and the area size of the trajectory anonymity region represents data utility. However, there is an inverse relationship between the data utility and the level of trajectory privacy. Since trajectory direction and similarity that is taken into consideration for a better trajectory privacy protection may increase the size of anonymity region, it may have negative effect on the accuracy of the results and will reduce the data utility. Therefore, according to the user's scenario requirements, we consider two types of proportions of trajectory privacy and data utility.

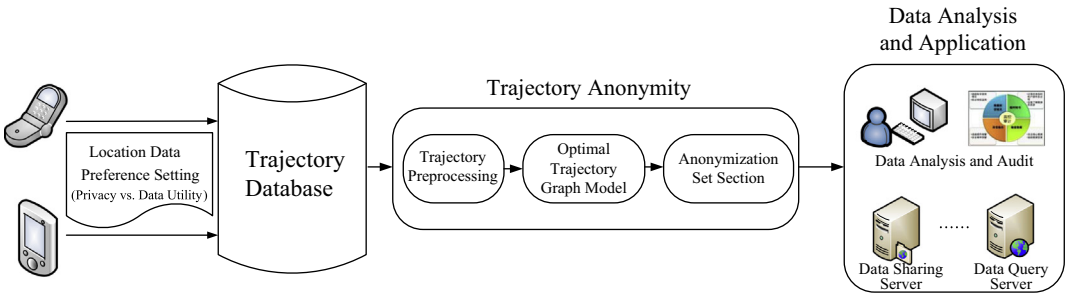

Fig. 2. A personalized trajectory anonymization model.

In the first case, it requires high data utility more than trajectory privacy. That is, it is desirable to find any k trajectories as candidates for anonymization, which can lead to a small size of anonymity region. In the second case, it mainly focuses on trajectory privacy. To prevent adversaries from identifying the user's real trajectory, the differences among k trajectories should not be too large. Most of the works on the basis of trajectory k-anonymity technique (Abul et al., 2008; Huo et al., 2011) measure the differences of trajectories and construct an anonymity region based on Euclidean distance. However, they may not consider the factor of users' movement directions. In this paper, we simplify our proposed metric slope ratio (Gao et al., 2012) as trajectory angle to measure trajectory similarity and construct an anonymity region on the basis of trajectory distance. According to the user's preference setting on the proportions of trajectory privacy and data utility, we construct a personalized anonymization model from the perspective of graph theory to find an optimal trajectory k-anonymity set.

We translate the relationship among trajectories into undirected weighted graph model by Definition 4.2. Algorithm 1 describes the process of trajectory graph construction in detail. The inputs of the algorithm are trajectory equivalence class and preference setting and the output is a trajectory graph $G=\{V,E,W\}$, where W presents correlation degree between trajectories.

Definition 4.2 (*Trajectory graph model*). A trajectory equivalence class can be converted into undirected weighted graph $G=\{V,E,W\}$, where vertexes V denote trajectories. There exists an edge between two vertexes V_i and V_j if they toward the same direction. The weight W_{ij} of the edge (T_i,T_j) represents the tradeoff of trajectory privacy and data utility between T_i and T_j.

Algorithm 1. Trajectory Graph Construction.

Require: Trajectory equivalence class $T=T_1\cup T_2\cup\cdots\cup T_n$, Preference setting α,β
Ensure: Trajectory graph $G=\{V,E,W\}$
$V\leftarrow T_1$, $E\leftarrow\Phi$, $W\leftarrow\Phi$;
$V'\leftarrow T-V$;
while $\sim IsEmpty(V')$ **do**
  for $T_i\in V$ & $T_j\in V'$ **do**
    $W_{ij}\leftarrow$ Weight Construction Process (T_i,T_j,α,β);
    $W\leftarrow W_{ij}$, $E\leftarrow(T_i,T_j,W_{ij})$;
    $V\leftarrow V+T_j$, $V'\leftarrow T-V$;
  end for
end while
Return $G=\{V,E,W\}$;

Most of the previous works have concentrated on achieving anonymity based on the aforementioned three types of methods by considering trajectory as a quasi-identifier. However, the differences of movement directions may facilitate each trajectory to be identified easily. As the two types of requirements we mentioned, there are conflicts between trajectory privacy and data utility. Based on the request of trajectory privacy α and data utility β, we establish weight function to balance the relationship of requirements.

Definition 4.3 (*Weight function*). Given a trajectory equivalence class with n trajectories $T=\{T_1,T_2,\ldots,T_n\}$. Let $S=\{S_1,S_2,\ldots,S_n\}$ refer to the trajectory similarity and $D=\{D_1,D_2,\ldots,D_n\}$ be the trajectory distances, where S_i and D_i are n-dimensional vectors and each dimension respectively represents the corresponding cosine sum of trajectory angles and trajectory distance between T_i and T_j, $j=1,2,\cdots,n$. After normalization, with the trajectory privacy and data utility configuration (α,β), the weighted tuple, denoted by $W=\{W_1,W_2,\ldots,W_n\}$, is defined as (5).

$$W_i=\alpha\cdot(1-S_i)+\beta\cdot D_i,\quad i=1,2,\ldots,n \tag{5}$$

where S_i represents the set of normalized similarity among trajectories; data utility is measured by anonymity regions which are inversely proportional to the normalized trajectory distances D_i. α and β indicate the demand preference which are assigned by a user and satisfy $\alpha,\beta\in[0,1]$ and $\alpha+\beta=1$.

Based on the weight function, Algorithm 2 summarizes the weight construction process. The inputs are any two different trajectories and preference setting and the output is the corresponding weight.

Algorithm 2. Weight Construction Process.

Require: Two Trajectories $T_r,T_s\in T(r\neq s)$, Preference setting α,β
Ensure: Weight W_{rs}
$S_{rs}\leftarrow 0$, $D_{rs}\leftarrow Distance(T_r,T_s)$, $i\leftarrow 1$;
while $i<n$ **do**
  for each trajectory segment $T_r^i\in T_r$, $T_s^i\in T_s$ **do**
    $S_{rs}^i\leftarrow\cos(T_r^i,T_s^i)$;
    if $S_{rs}^i\leq 0$ **then**
      $S_{rs}^i\leftarrow 0$;
    end if
    $S_{rs}\leftarrow S_{rs}+S_{rs}^i$;
  end for
  $i\leftarrow i+1$;
end while
Normalize trajectory similarity S_{rs} and distance D_{rs};
$W_{rs}\leftarrow\alpha\cdot(1-S_{rs})+\beta\cdot D_{rs}$;
Return W_{rs};

4.3. Anonymization set selection

The following problem is how to find a trajectory k-anonymity set from the constructed optimal trajectory graph model. The selection of the trajectory k-anonymity set affects the trajectory privacy level and data utility. Each user requires k similar trajectories with smaller size of anonymity region. That is to request S_{ij} to be larger and D_{ij} to be smaller enough. In other word, the weight W_{ij} of the selected k trajectories should be smaller as possible. Differ from the goal of Huo et al. (2011), they only considered the Euclidean distance for data utility and formalized the problem into graph partition problem. In this paper, the optimal k trajectories selection corresponds to a constrained minimum spanning tree problem.

$$W=\begin{array}{c} \\ T_1 \\ T_2 \\ \cdots \\ T_n \end{array}\begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \cdots & T_n \end{array} \\ \begin{pmatrix} 0 & W_{12} & \cdots & W_{1n} \\ W_{21} & 0 & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & 0 \end{pmatrix} \end{array}$$

The weight distribution W represents the requirements of privacy protection and data utility with the same α and β. Obviously, W is a symmetric matrix and the diagonal elements are equal to 0. That is, $\forall i,j=1,2,\ldots,n, W_{ij}=W_{ji}\wedge W_{ii}=0$. According to our analysis, regardless of a user's preference settings, it requests the selected k trajectories with a higher similarity S_{ij} and a smaller distance D_{ij}. Algorithm 3 depicts the initial search point selection and edge set. To select k vertexes with the smaller weights, we first search an edge with minimum weight and specify the two vertexes affiliated with the edge as the start point set. Then, the *GreedySearch* method is called to find other k–1

trajectories with smaller weights. In *GreedySearch*, we adopt a greedy strategy such as Prim's algorithm to search k trajectories as trajectory k-anonymity set K.

Algorithm 3. Initial Point Selection.

Require: Trajectory graph $G=\{V,E,W\}$, Requirement k
Ensure: $K=\{k$ trajectories$\}$
1: $V'\leftarrow\Phi$, $E'\leftarrow\Phi$, $K\leftarrow\Phi$;
2: minValue$\leftarrow W(1,1)$, $r\leftarrow 0$, $s\leftarrow 0$;
3: **if** $k<|V|$ **then**
4: **for** $i=1:|V|$ **do**
5: **for** $j=1:|V|$ **do**
6: **if** $j\neq i$ and minValue $> W(i,j)$ **then**
7: minValue$\leftarrow W(i,j)$;
8: $r\leftarrow i, s\leftarrow j$;
9: **end if**
10: **end for**
11: **end for**
12: $V'\leftarrow\{v_r,v_s\}$, $E'\leftarrow\{(v_r,v_s)\}$;
13: $W(r,s)\leftarrow 0$, $W(s,r)\leftarrow 0$;
14: $K\leftarrow$ GreedySearch (G,V',E',k);
15: **else**
16: k is not available;
17: **end if**

Greedy algorithm such as Prim's algorithm (Greedy algorithm) is essentially a heuristic method that yields locally optimal solutions at each stage to approximate a global optimal solution. In this paper, we search a relatively optimal anonymity set K based on Algorithm 4. The main idea of this algorithm is to find the minimum edge weight to the other connected edge weights with the number of vertexes is less than k. The algorithm starts from the vertex affiliated with another vertex with minimum edge weight that is computed by Algorithm 3. Two group sets V' and E' represent the vertexes that have been selected and their adjacent edge set. While the number of selected vertexes is less than k, V' spreads itself by searching the vertexes that connect it with the minimum edge weight to the other edge weights, and then adding that vertex into V' and the corresponding edge into E'. Once a vertex has been added to V', we set the adjacent edge weight to 0. The algorithm loops until it finds k vertexes that meet the condition.

4.4. A case study

To demonstrate these algorithms clearly, we run a simple example to describe how $k=5$ trajectories are selected. Figure 3 (a) represents the original graph constructed by trajectories with the weight matrix denote by W as follows:

$$W=\begin{array}{c} \\ T_1\\T_2\\T_3\\T_4\\T_5\\T_6\\T_7\end{array}\begin{array}{c}\begin{array}{ccccccc}T_1&T_2&T_3&T_4&T_5&T_6&T_7\end{array}\\\begin{pmatrix}0&10&6&5&5&0&7\\10&0&7&3&0&7&6\\6&7&0&0&1&3&4\\5&3&0&0&2&5&4\\5&0&1&2&0&4&5\\0&7&3&5&4&0&2\\7&6&4&4&5&2&0\end{pmatrix}\end{array}$$

According to Algorithm 3, we search the start vertex set $V'=\{v_3,v_5\}$ and $E'=\{(v_3,v_5)\}$. Algorithm 4 finds the adjacent vertex v_4 to $V'=\{v_3,v_5\}$ with the minimum weight. The vertex is added to $V'=\{v_3,v_5,v_4\}$ and the edge is added to $E'=\{(v_3,v_5),(v_5,v_4)\}$. While $|V|<k$, it repeats the process until finding the other vertexes and edges that satisfy the condition. Figure 3 (b) shows an anonymity set $K=V'=\{v_3,v_5,v_4,v_6,v_7\}$ is constructed that meet $k=5$ trajectories with minimum weight.

Algorithm 4. GreedySearch(G,V',E',k).

Require: Graph G, Requirement k, Start Point V' and Edge E'
Ensure: $K=\{k$ trajectories$\}$
1: $V'\leftarrow\{v_r,v_s\}$, $E'\leftarrow\{(v_r,v_s)\}$;
2: **while** $|V'|<k$ **do**
3: **for** $i=1:|V'|$ **do**
4: $temp\leftarrow W(V'(i),:)>0$; //filter selected vertex
5: $minWeight(i)\leftarrow min(W(V'(i),temp))$; //find minimum edge weight of each selected vertex
6: $index\leftarrow find(W(V'(i),:)==minWeight(i))$;
7: $index\leftarrow index(1)$; // take the first index of the minimum vertex
8: $minValue(i,:)\leftarrow[V'(i),index,minWeight(i)]$;
9: **end for**
10: $[a,b]\leftarrow min(minValue(:,3))$; //find the minimum weight and corresponding index
11: $V'\leftarrow\{V',minValue(b,2)\}$; //add to V'
STATE $E'\leftarrow$Add the corresponding edge; //add to E'
//delete the selected vertex
12: $W(V',minValue(b,2))\leftarrow 0$;
13: $W(minValue(b,2),V')\leftarrow 0$;
14: **end while**
15: Return $K\leftarrow V'$;

5. Privacy and data utility

The correlation degree among trajectories is quantified by weight function with considering the factors of trajectory privacy and data

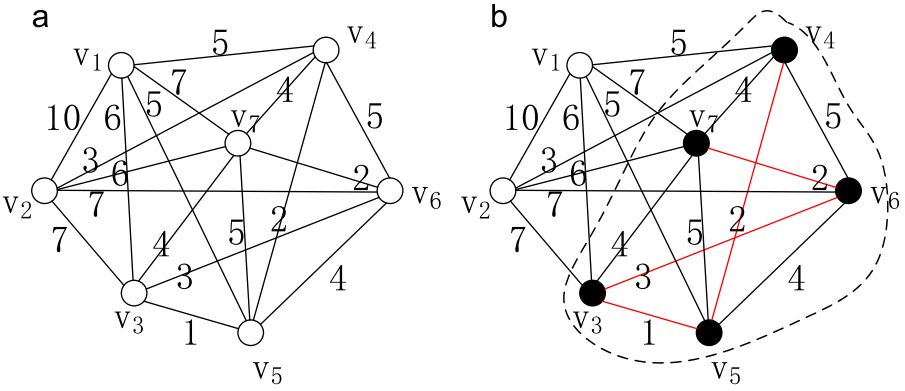

Fig. 3. A simple example. (a) Original graph and (b) search k trajectories.

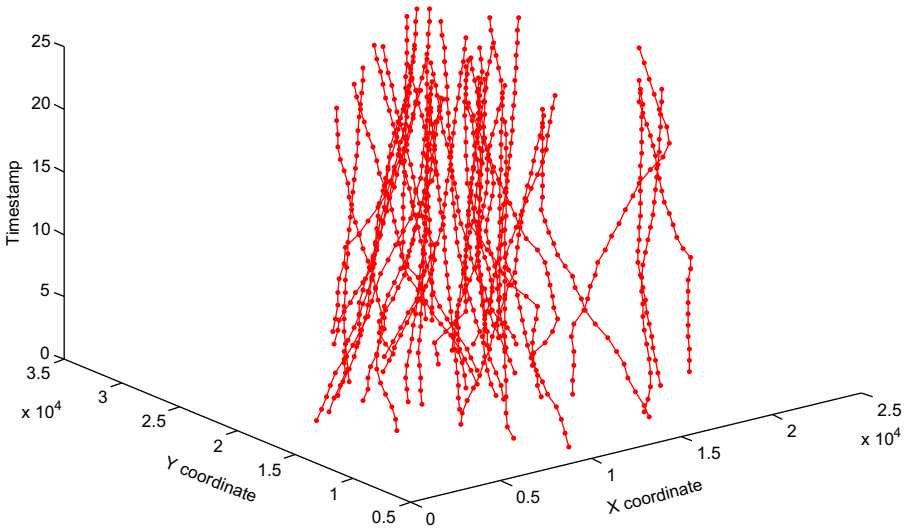

Fig. 4. An equivalence class.

utility. The selection of a trajectory k-anonymity set is dependent on the distributions of weights. In this section, we introduce trajectory similarity to evaluate the effectiveness of the selected k trajectories, and then analyze information loss quantitatively.

5.1. Privacy level

Privacy protection is always measured by the probability of adversaries to identify a trajectory from published database. The GreedySearch algorithm achieves trajectory k-anonymity based on the selected k vertexes from graph G with minimum weight. Considering different proportions of privacy and data utility requirements, we provide different optimal k-anonymity sets for trajectory privacy protection. Because we take the similarity and direction between trajectories into account, compared with previous work, we can provide a personalized optimal trajectory k-anonymity set.

Theorem 5.1. *Given a trajectory equivalence class* $T=\{T_1, T_2, T_3, \ldots, T_n\}$. *We form a trajectory k-anonymity set* $K=\{T_1', T_2', \ldots, T_k'\}$ *by finding optimal k similar trajectories with minimum information loss based on GreedySearch. The average probability of identifying the trajectory is bounded by* $1/k$.

Proof. We assume that an adversary can obtain the selected trajectory k-anonymity set and public knowledge. In the worse condition, the adversary has not aware of the privacy protection model, the probability of privacy invasions is reduced to $1/area$, where *area* represents the area size of anonymity area. Due to the similarity among trajectories, in the best case even if the adversary knows the size of the trajectory k-anonymity set and those sampling locations on each trajectory that anonymized together, the probability of privacy leak is under $1/k$. □

In addition to traditionally using the number of trajectories k to measure the privacy in Theorem 5.1, in this paper, we evaluate the privacy level by analyzing the relationship among trajectories. That is to evaluate the privacy level only by the similarity of the selected k trajectories. Let the selected k trajectories be $K=\{T'_1, T'_2, \ldots, T'_k\}$, where each trajectory is represented as n spatiotemporal points according to Definition 3.1. The average similarity of the trajectory k-anonymity set, denoted as S_{avg}, can be computed by (6).

$$S_{avg}=\frac{\sum_{i=1}^{k}\sum_{j=1\wedge j\neq i}^{k} S(T'_i, T'_j)}{k\cdot(k-1)} \tag{6}$$

The privacy level, denoted as PL, is defined as the ratio of average similarity of the trajectory k-anonymity set to that of the maximum one, which is depicted by (7). In extreme condition, each trajectory segment of the two trajectories is fully parallel in the same direction. That is, the cosine of each trajectory segment between T'_i and T'_j is equal to 1, thus, S_{avg} reaches the maximum similarity $n-1$.

$$PL=\frac{S_{avg}}{n-1} \tag{7}$$

5.2. Data utility

Follow by our analysis above, data utility is measured by anonymity regions, which is inversely proportional to information loss. Information loss is mainly caused by generalizing trajectories to a region, which makes data utility degenerate. Thus, in information loss evaluation, we only account for the generalization part, while the information loss caused by pre-processing phase does not take into consideration. In Yarovoy et al. (2009) and Huo et al. (2012), they adopt the reduction in the probability with which people can accurately determine the position of an object. In this paper, we measure the information loss by the ratio of the size of trajectory k-anonymity region to the area size of the whole space, which is computed by (8).

$$IL=\frac{Area(O, D'_{max})}{MaxArea} \tag{8}$$

where $Area(O, D'_{max})$ represents the area size of generalized region of the trajectory k-anonymity set with the diameter D'_{max} and $MaxArea$ represents the size of the whole space.

As demonstrated by the procedure of the anonymity region construction in Section 3.1, the anonymity region R is formed by the circular area with maximum trajectory distance. When the area size of the trajectory k-anonymity region approximates to the size of the whole space $MaxArea$, the information loss is close to the maximum. Since data utility is inversely proportional to the information loss, at this point, we consider that the data utility reduces to the minimum.

6. Evaluation

In this section, we report the evaluation results we have conducted in order to assess the performance of our method, in terms of privacy protection level that our method can achieve and

data utility maintained in the anonymity process. In particular, we are interested in the performance under various proportions of requirements on privacy protection and data utility.

In short, our experimental evaluation consists of two parts: (1) According to different preference settings, we evaluate the effectiveness of our method and compare it with previous work in terms of trajectory privacy and data utility; (2) We compare the efficiency of our trajectory k-anonymity model with previous work.

6.1. Experiment set

In all the experiments, we use Thomas Brinkhoff Network-based Generator of Moving Objects (Brinkhoff, 2003) to generate a set of moving objects. The input to the generator is the road map of Oldenburg in Germany with an area of about 200 km^2. The outputs are 100,000 trajectories that describe the movement of objects within one day along the road-network of the city Oldenburg. After the phase of trajectory pre-processing, we randomly select a trajectory equivalence class with 40 trajectories depicted by Fig. 4. Each trajectory in the equivalence class is represented as a sequence of 26 spatiotemporal points, that is $n=26$.

6.2. Effectiveness

Under different proportions of trajectory privacy and data utility requirements, recall the goal of our trajectory anonymity is to find a personalized optimized trajectory k-anonymity set with minimum cost. We evaluate the effectiveness of our method in terms of trajectory privacy protection and data utility.

6.2.1. Trajectory privacy protection

Based on the privacy metric in Section 5.1, we utilize trajectory similarity to evaluate trajectory privacy level. Adversaries can hardly distinguish trajectories easily if trajectories are similar. We observe that trajectory direction affects the quality of anonymity set. We introduce trajectory angle to measure trajectory similarity and direction and add it into the process of trajectory k-anonymity set selection. In this section, we first measure the privacy level of the selected trajectory k-anonymity set can achieve in different preference settings and then compare it with previous works (Abul et al., 2008; Huo et al., 2011).

From the perspective of a user's requirements, Fig. 5 presents different proportions of preference settings on trajectory protection and data utility. We set four groups of different user's preference settings on trajectory privacy and data utility: $(\alpha,\beta)=\{(0,1),\ (0.4,0.6),(0.8,0.2),(1,0)\}$. The selection of a trajectory k-anonymity set depends on the specific application scenarios and the user's requirements. There are two special cases: (1) Only focus on data utility $(\alpha,\beta)=(0,1)$. For example, if an emergency incident happens suddenly in the anonymity process, the first reaction is how to let himself/herself be discovered as possible. It means that the user only concerns on data utility. In this case, it is degenerated to the works such as Abul et al. (2008) and Huo et al. (2011) that only consider data utility without trajectory similarity and direction to select the trajectory k-anonymity set; (2) Only concern with privacy level $(\alpha,\beta)=(1,0)$. Recently, Google[1] and Apple[2] were trapped in privacy laws for tracking users' locations. In this case, users just concern on serious risk of privacy invasions. Besides, in most cases, the users have different emphases respectively on the aspect of trajectory privacy and data utility. Thus, according to each user's requirements, he/she requires different degrees of privacy level and data utility.

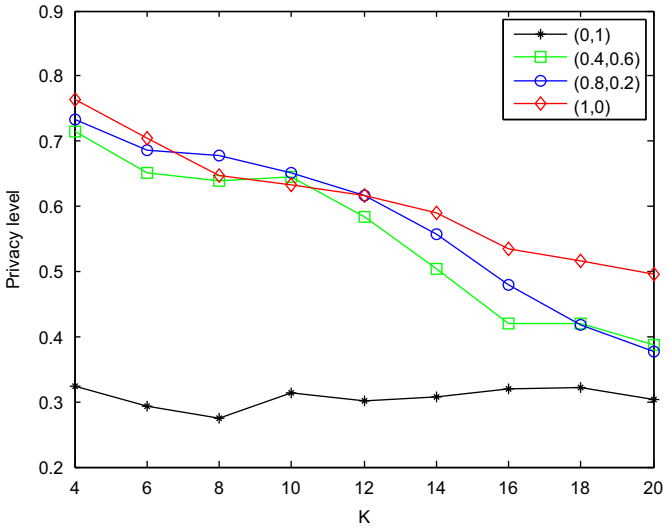

Fig. 5. Privacy level evaluation.

As illustrated by Fig. 5, as expected, the privacy level is higher with more concern on trajectory similarity and direction. That is a larger value of α will cause a preferable trajectory k-anonymity set. Previous works (Abul et al., 2008; Huo et al., 2011) only focus on data utility, that is $(\alpha,\beta)=(0,1)$, which cause a low privacy level. Note that we take the average similarity of the selected trajectory k-anonymity set as the evaluation criteria to quantitatively analyze privacy level. On the whole, there is a decrease in privacy level with the number of trajectories k increases in the same preference setting. That is because with the increase of the size of trajectory k-anonymity set, it is more difficult to find more similar trajectories. Instead, a smaller size of trajectory k-anonymity set might make it much easier to find k similar trajectories.

6.2.2. Data utility

On the contrary, a higher trajectory privacy level would reduce the data utility. With different trajectory k-anonymity sets, data utility is inversely proportional to the area size of the corresponding anonymity region. Similarly, we consider the same preference setting above. We have analyzed the main cause of the information loss and presented the metric in Section 5.2. Figure 6 shows the results of information loss under various allocation proportions. Overall, the information loss grows with the number of trajectories increases under the same proportion. Since a larger value of k may cause a larger area size of the generalized region, it results in more information loss. Besides, a more attention on data utility causes less information loss than others. That is because the requirements on data utility affects the selection of the trajectory k-anonymity set. Specifically, the works (Abul et al., 2008; Huo et al., 2011) cause the minimum information loss with the selected trajectory k-anonymity set. That is because they concern more about data utility in the anonymity process with $(\alpha,\beta)=(0,1)$. A higher requirement on data utility gets the trajectory k-anonymity set with a smaller anonymity region, which causes a lower information loss. It can be seen that, when k reaches a certain value, the growth of information loss is relatively stable, meaning that the increasing of the size of trajectory k-anonymity set will not sharply increase the information loss.

6.3. Efficiency

We compare the efficiency of our method with the other two cases: In the trajectory anonymity process, one is paid more

[1] Mills E. Google sued over android data location collection. http://news.cnet.com/8301-27080_3-20058493-245.html, April 2011, CNET News.

[2] Lowensohn J. Apple sued over location tracking in iOS. http://news.cnet.com/8301-27076_3-20057245-248.html, April 2011, CNET News.

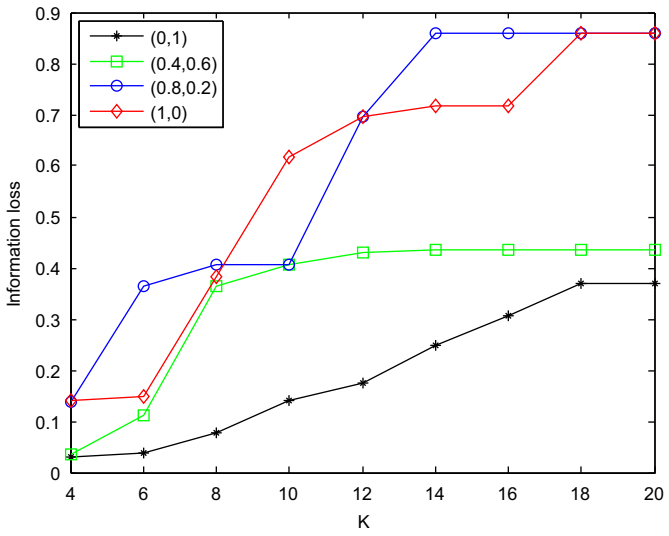

Fig. 6. Information loss evaluation.

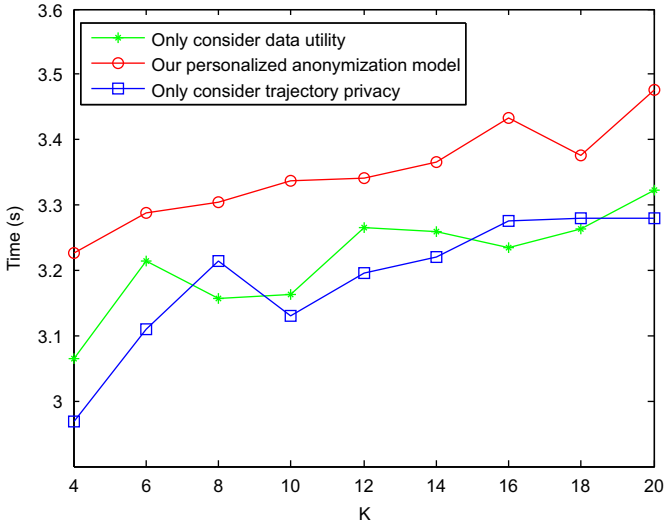

Fig. 7. Run-time comparison.

attention on data utility such as Huo et al. (2011) and the other is more concerned about trajectory privacy on a Intel(R) Core(TM) 2 Quad 2.83 GHz processor with 4 GB RAM on Windows XP platform.

We run several times to get the average execute time to evaluate the efficiency of our method. Figure 7 depicts the run-time comparison of our method with the other cases. As Fig. 7 illustrates, the execute time of our method and Huo et al. (2011) increases with the growth of trajectory k-anonymity set. That is because the growth of k needs more time to find the optimal trajectory k-anonymity set that satisfies the requirements. Compared our method with the other cases, we can find that the time of our method required is a little more than they need. That is because our method takes the factors of trajectory similarity and data utility all into consideration, while the other two cases only concern one of them in the anonymity process. Therefore, according to the weight construction process, our method needs more time than the other cases to search an optimal trajectory k-anonymity set. Overall, we can find a more preferable trajectory k-anonymity set to balance trajectory privacy protection and data utility at the cost of reducing a little efficiency.

7. Conclusion and future work

This paper concerns a personalized anonymization model to balance trajectory privacy and data utility. Most of trajectory k-anonymity methods ignore trajectory similarity and direction in the anonymity process. However, adversaries may use the difference of anonymity trajectories to identify each trajectory. Meanwhile, the data utility of trajectory may reduce with the expansion of an anonymity region. Motivated by this, this paper takes all the factors into account. We propose to use trajectory angle to evaluate trajectory similarity and direction and construct the anonymity region based on trajectory distance. Considering the various proportional distributions of trajectory privacy and data utility requirements in different scenarios, we propose a personalized anonymization model to select a trajectory k-anonymity set. Compared with prior work, it demonstrates that our model can provide an optimal trajectory k-anonymity set by analyzing the effectiveness in terms of privacy level and data utility. Meanwhile, it has a little impact on efficiency.

Our further work is to deploy the prototype system that integrates our personalized anonymization model and develop the query processing that supports our model.

Acknowledgment

The authors appreciate the helpful comments and suggestions from the anonymous referees. This work was supported by Program for Changjiang Scholars and Innovative Research Team in University (IRT1078), the Key Program of NSFC-Guangdong Union Foundation (U1135002), Major National S&T Program (2011ZX03005-002), the Grants from the Natural Science Foundation of China (61072066) and the Fundamental Research Funds for the Central Universities (JY10000903001, K5051203010).

References

Abul O, Bonchi F, Nanni M. Never walk alone: uncertainty for anonymity in moving objects databases. In: IEEE 24th international conference on data engineering (ICDE 2008). IEEE; 2008. p. 376–85.
Brinkhoff T. Generating traffic data. IEEE Data Engineering Bulletin 2003;26(2):19–25.
Cao X, Cong G, Jensen CS. Mining significant semantic locations from GPS data. Proceedings of the VLDB Endowment 2010;3(1–2):1009–20.
Chen L, Özsu MT, Oria V. Robust and fast similarity search for moving object trajectories. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. ACM; 2005. p. 491–502.
Domingo-Ferrer J, Trujillo-Rasua R. Microaggregation- and permutation-based anonymization of movement data. Information Sciences: An International Journal 2012;208:55–80.
Gao S, Ma J, Shi W, Zhan G. Towards location and trajectory privacy protection in participatory sensing. In: Proceedings of the 3rd International Conference on Mobile Computing, Applications, and Services; Log Angles, USA, 2011. p.381–386.
Gedik B, Liu L. Location privacy in mobile systems: a personalized anonymization model. In: Proceedings of the 25th IEEE international conference on distributed computing systems (ICDCS2005). IEEE; 2005. p. 620–9.
Greedy algorithm, ⟨http://en.wikipedia.org/wiki/Greedy_algorithm⟩.
Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the first international conference on mobile systems, applications and services (MobiSys2003). ACM; 2003. p. 31–42.
Gruteser M, Liu X. Protecting privacy in continuous location-tracking applications. IEEE Security & Privacy 2004;2(2):28–34.
Huo Z, Huang Y, Meng X. History trajectory privacy-preserving through graph partition. In: Proceedings of the first international workshop on mobile location-based service. ACM; 2011. p. 71–8.
Huo Z, Meng X, Hu H, Huang Y. You can walk alone: trajectory privacy-preserving through significant stays protection. In: Proceedings of the 17th international conference on database systems for advanced applications (DASFAA2012); 2012. p. 351–366.
Ivanov R. Real-time GPS track simplification algorithm for outdoor navigation of visually impaired. Journal of Network and Computer Applications 2012:1559–67.

Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services. In: Proceedings of international conference on pervasive services (ICPS'05). IEEE; 2005. p. 88–97.

Nergiz ME, Atzori M, Saygin Y. Towards trajectory anonymization: a generalization-based approach. In: Proceedings of the SIGSPATIAL ACM GIS 2008 international workshop on security and privacy in GIS and LBS. ACM; 2008. p. 52–61.

Pelekis N, Kopanakis I, Marketos G, Ntoutsi I, Andrienko G, Theodoridis Y. Similarity search in trajectory databases. In: 14th international symposium on temporal representation and reasoning. IEEE; 2007. p. 129–40.

Shin H, Vaidya J, Atluri V. Anonymization models for directional location based service environments. Computers & Security 2010a;29(1):59–73.

Shin H, Vaidya J, Atluri V, Choi S. Ensuring privacy and security for lbs through trajectory partitioning. In: 2010 11th international conference on mobile data management (MDM'10). IEEE; 2010. pp. 224–6.

Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories. In: Ninth international conference on mobile data management (MDM'08). IEEE; 2008. p. 65–72.

Tiakas E, Papadopoulos AN, Nanopoulos A, Manolopoulos Y, Stojanovic D, Djordjevic-Kajan S. Searching for similar trajectories in spatial networks. Journal of Systems and Software 2009;82(5):772–88.

Trajcevski G, Wolfson O, Hinrichs K, Chamberlain S. Managing uncertainty in moving objects databases. ACM Transactions on Database Systems (TODS) 2004;29(3):463–507.

Xiao Z, Meng X, Xu J. Quality aware privacy protection for location-based services. In: Proceedings of the 12th international conference on database systems for advanced applications (DASFAA2007). Springer-Verlag; 2007. p. 434–46.

Xu T, Cai Y. Exploring historical location data for anonymity preservation in location-based services. In: The 27th conference on computer communications (INFOCOM). IEEE; 2008. p. 547–55.

Yarovoy R, Bonchi F, Lakshmanan LVS, Wang WH. Anonymizing moving objects: how to hide a mob in a crowd? In: Proceedings of the 12th international conference on extending database technology: advances in database technology. ACM; 2009. p. 72–83.

You TH, Peng WC, Lee WC. Protecting moving trajectories with dummies. In: 8th international conference on mobile data management (MDM'07), IEEE; 2007. p. 278–82.

Zheng Y, Zhang L, Xie X, Ma WY. Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th international conference on world wide web (WWW2009). ACM; 2009, p. 791–800.